ボイスダイヤリングシステムの現状

河井　恒　　樋口宜男
(株) KDD研究所

●身近になってきたボイスダイヤリング

最近，ボタンを操作する代わりに音声で相手の名前を言うと電話がかけられる携帯電話端末が売られているのをご存知だろうか．歩きながら電話をかけるときなど，いちいちボタンを見ていられない場面では重宝する．この機能は，自動音声認識技術のささやかな応用であり，ボイスダイヤリング（以下，VD）と呼ばれるものである．本稿では，このボイスダイヤリングについて解説する．

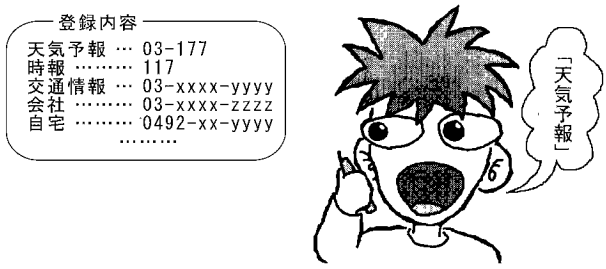


図-1　ボイスダイヤリングの使い方

●ボイスダイヤリングの使い方

VD機能付き電話端末の一般的な使い方を説明しておこう（図-1）．03-177に電話をかけるときに，番号をダイヤルする代わりに“テンキヨホウ”と発声したいとする．

そのためには，まず音声と電話番号を電話端末に登録しておかなければならない．最初のステップとして，VDを登録モードにする．次に，電話番号（03-177）を入力する．次に電話端末からのガイダンスにしたがって“テンキヨホウ”と発声する．この音声は，電話番号に付したラベルという意味で音声ラベルと呼ばれることがある．同じ音声ラベルを2，3回発声するよう求められることがあるが，これは言いよどみや雑音の影響を排除して登録を確実に行うためである．最後に，電話端末から電話番号と音声ラベルが再生され，確認を求められるので，内容が正しければこれで1件分の登録が終了する．さらにほかの相手先を登録したい場合は，同じ操作を繰り返す．現在国内で市販されている製品では，登録できる件数は5～20件である．

電話をかける場合は，まずVDを発呼モードにする．次に電話端末に向かって音声ラベル（“テンキヨホウ”）を発声する．音声認識処理が終了すると「“テンキヨホウ”におつなぎします」というようなメッセージが流れるので，認識結果が正しければそのまま接続を待ち，誤っていれば始めからやり直す．

●ボイスダイヤリングを分類する

ここでは，VDを「音声によって電話の相手先を指定する機能」と解釈した上で，発声内容，登録手段・登録者，音声認識アルゴリズム，実装場所，の4つの観点から分類する．**表-1**にそれぞれの観点についての選択肢の現実的な組合せを示した．ただし，実装場所は，表-1の4通りの組合せすべてについてセンタと端末の2通りが可能であるので，表では省略した．

発声内容には，相手先の電話番号自体と音声ラベルの2通りがある．電話番号を発声する場合は，登録操作は不要であるが，音声ラベルの場合は，サービスの利用者または提供者が事前に登録操作を行う必要がある．サービス提供者が登録を行う場合，登録内容は公共機関，知名度の高い会社名などになる．このようなシステムは，ボイスディレクトリ，音声フリーフォンなどと呼ばれることもある．

利用者が登録を行う手段としては，音声と文字の2通りがある．音声で登録した場合は，登録された音声を再生することによって利用者に認識結果の確認を求めることができるが，文字の場合には，それができない．このため，多少の不自然さに目をつぶって合成音声を使うか，あるいは別途録音した音声を使用する．

発声内容	登録手段	登録者	認識アルゴリズム	サービスの形態
電話番号	(登録不要)		HMM	電話番号発声によるボイスダイヤリング
音声ラベル	音声	提供者	HMM	ボイスディレクトリ，音声フリーフォン
		利用者	HMM，DTW	狭義のボイスダイヤリング
	文字	利用者	HMM	

表-1　ボイスダイヤリングの分類

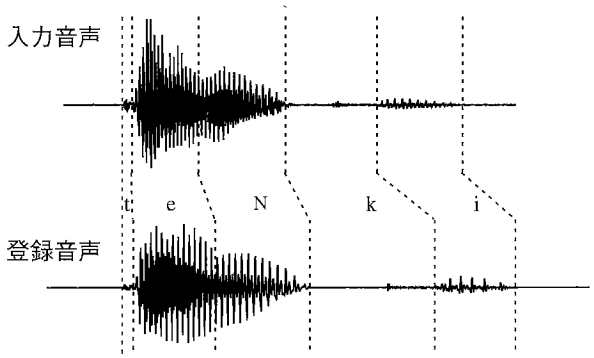


図-2　DTWによる音声認識．入力音声の時間軸を非線型に伸縮して参照音声の時間軸と対応づけた上で両者の類似度を計算する．言語によらず認識できる，実装が簡単である，という利点がある．

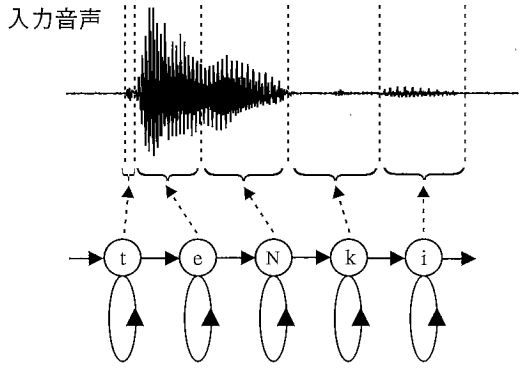


図-3　HMMによる音声認識．入力音声を生成した確率が最も高い状態系列とその確率値を推定する．不特定話者用の音素HMMを使えるので，事前登録が可能である．

●音声認識には2つの手法がある

VDの音声認識処理は，基本的には登録音声と入力音声のパターン照合である．具体的なアルゴリズムとしては，DTW（Dynamic Time Warping，動的時間正規化）またはHMM（Hidden Markov Model，隠れマルコフモデル）が用いられる．DTWは，登録音声と入力音声を直接照合する．HMMは，確率の概念を持ち込むことによって，発話速度，音素☆1環境，話者性などさまざまな揺らぎを織り込みつつパターン照合を行うことができる．

◎言語によらず認識できるDTW

音声の発話速度は，発声ごとに微妙に変動し，しかも変動の程度は音素ごとに異なる．このため，登録音声と入力音声を照合する際には，時間軸を線形に伸縮しただけでは正しい類似度を求めることができない．DTWでは，DP（Dynamic Programming，動的計画法）アルゴリズムによって時間軸を非線型に伸縮し，入力音声と登録音声の類似度を求める（図-2）．

DTWの長所は，アルゴリズムが簡単であり，必要となる計算量，メモリともに比較的少ないことである．また，言語依存性がないため，同一の装置を日本語にも英語あるいはそれ以外の言語にも使うことができる．ただし，日本語の音声認識では，通常，声の高さを考慮しないため，「箸」と「橋」のようにアクセントのみが異なる単語は区別できない．これは次に紹介するHMMでも同様である．短所は，話者あるいは音素環境に起因する音響的特徴の揺らぎをうまく吸収できないため，不特定話者の音声認識に適さないことである．

これらの特徴のために，現在ではもっぱら数十語以下の孤立単語の特定話者認識に使われる．ただし，特定話者とはいっても，異なる話者の音声を正しく認識できる場合が少なくない．

◎登録が不要なHMM

HMMでは，登録音声を音素列の形で記憶する．登録内容を入力音声と照合する際には，個々の音素に対するHMMを接続して状態遷移ネットワークを作成し，このネットワークから入力音声が生成されたと仮定した場合の確率を計算する（図-3）．この確率が，登録音声と入力音声の類似度に相当する．音素HMMから音声信号が出力される際の確率分布は，多様な話者および発声内容を含む大量の音声データを用いて学習する．

HMMをVDに適用する場合，文字による登録では，登録内容を音素HMMの列に変換し，単語認識を行う．通

☆1 言語音声を構成する基本的な単位．言語ごとに異なり，日本語の場合は5母音ほか全部で40個程度ある．

認識アルゴリズム	DTW
音声検出方式	パワーによる
発声内容	任意の語句
登録語数	最大10語
音声長	発声あたり最大3秒
回線種別（装置側）	ISDN（BRI）
CPU	HyperSPARC-125MHz

表-2　フィールド試験で用いたVDシステムの仕様

対象	回線	人数	呼数	認識率(%)
社員	加入者線	172	6426	98.2
	公衆電話	115	1879	95.3
一般	加入者線	176	3766	99.5
	公衆電話	23	245	98.4

表-3　VDフィールド試験の結果

常は不特定話者用の音素HMMを使うので，話者依存性はない．

これに対して，音声による登録の場合は，まず音素認識を行い，得られた音素列を電話帳に登録する．現在の音素認識の精度は60～70%程度であるため，登録される音素列は正しい表記とは異なるのが普通である．両者の差分には個人性が含まれると考えられることから，基本的に特定話者認識となる．

◎時代の流れはHMM？

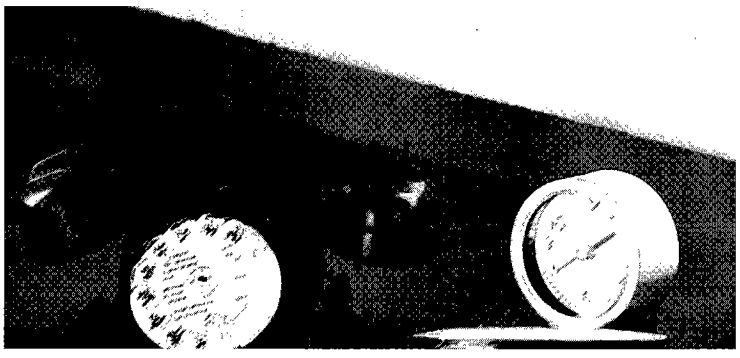
利用者が相手先を登録する形の一般的なVDでは，DTWとHMMの認識性能は同等と考えてよい．

HMMには言語依存性があるが，言語が異なってもまったく認識できないわけではないし，音素モデルを差し替えるなど認識装置での対策も可能なので，実用上大きな問題とはなるまい．また，HMMはDTWと比較してアルゴリズムが複雑なため実装が難しいが，最近は組み込み用CPUの性能も向上してきたため，携帯電話への実装も可能となった．

HMMの大きな特長は，不特定話者認識が可能なことである．この特長から，サービス提供者が電話の相手先をあらかじめ登録しておくことができる，文字による登録も可能である，という利点が生じ，電話番号によるVD，音声フリーフォンなど多様なサービスが可能となる．

さらに，HMMの実装上の利点として，登録のための記憶容量が少ない点がある．具体的には，認識結果確認用の音声を考慮しない場合，記憶容量はDTWの1/300にすぎない．

これらのことより，今後，VDの認識アルゴリズムは，HMMが主流となっていくものと予想される．



●ボイスダイヤリングを実装する場所

VDを実装する場所としては，電話局内と端末内の2通りある．ここでは，前者をセンタ型，後者を端末型と呼ぶ．

センタ型の利点は，端末を選ばないことである．このため，家庭またはオフィスで登録して外出先から利用するという使い方もできるが，端末・回線を経由した音声を認識しなければならないという欠点を有する．特に携帯電話では，非線型な符号化のために認識率が低下する．伝送エラーも認識率低下の原因となる．なお，筆者らの実験によれば，PHSでは有線電話に匹敵する認識性能が得られることが確認されている．また，通信事業者は多数のVD利用者に対して均一のサービスを安定的に提供する義務があるので，VDサービスの開始・仕様変更・廃止には多大なコストが必要とされることも欠点といえよう．

一方，端末型の利点は，端末の周波数特性が既知であることと，伝送系による音声品質の劣化がないことである．また，端末型では基本的に個別の製品の購入者のみをサポートすればよいので，VDの仕様変更は比較的容易であろう．欠点は，端末ごとに登録を行わなければならない点である．ただし，携帯電話の普及にともなって出先でも同じ端末を使う機会が増えていることから，今後は問題とならなくなると思われる．

●国内でもセンタ型ボイスダイヤリングのフィールド試験が実施された

KDDでは，社員および一般加入者を対象としてセンタ型ボイスダイヤリングのフィールド試験を実施した[1]．システムの仕様を**表-2**に，結果を**表-3**に示す．公衆電話で認識率が若干低下する傾向があるものの，おおむね実用に耐える性能が得られている．平均的な応答時間（発声終了から認識結果が出るまでの時間）は，1.5秒程度であった．

これとは別に，シミュレーション実験ではあるが，電話番号をそのまま発声する実験も行っており，桁を区切ら

事業者	発売国	商品名	種別	備考
AT&T	米国	VoiceLine	センタ型	1995年開始
Bell Atlantic-Nynex Mobile	米国	TalkDial	センタ型	1996年開始
Sprint	米国	Total Voice	センタ型	1996年開始
KDD	日本	ボイスダイヤリング	センタ型	1996年試行
Korea Telecom	韓国	Voice-Dialing	センタ型	ボイスディレクトリ．1998年試行
Digital Acoustics	米国	Tell A Phone	端末型	PC用ソフト
Microsoft	米国	Cordless Phone	端末型	コードレス電話機
VCS	米国	VoiceDialer	端末型	電話機用アダプタ
ユニデン	日本	ダイヤレスホン	端末型	ハンズフリーマイク入力のみ
NECほか	日本		端末型	携帯電話各社

表-4　ボイスダイヤリングサービスおよび製品

ずに発声された（ただし，短い区切りを入れることは任意）10桁の数字に対して97.3%の認識率が得られている[2]．応答時間は，PentiumII™-333MHzベースのPCで約1秒である．実環境においては，数%の認識率の低下が予想されるが，それを割り引いてもおおむね実用段階に達しているといえよう．

●国内外のボイスダイヤリング製品

国内外のVDサービスおよびVD装置の実例を表-4にまとめた．

センタ型のVDは，早くからAT&T，NYNEX，SPRINTなど米国の通信事業者によって行われてきた[3]．国内では，1996年にKDDが一部の顧客を対象として試行サービスを行った[1]．Korea Telecomのサービス[4]は，表中の他社のサービスと異なり，事業者が500程度の組織（政府機関，報道機関，病院など）の音声ラベルを事前に登録しておく方式である．

端末型のVD製品は，携帯電話機，コードレス電話機，通常の電話機に外付けするアダプタなどさまざまな形態がある．ちなみに，携帯電話機の場合，認識率は静粛環境下で95%程度と報告されている[5]．

表-4に記載したもの以外にも，Lucent Technologies社やSensory Voice Activation社のように通信事業者やメーカに対して音声認識技術を供与している会社もある．

●ボイスダイヤリングは定着するか？

すでに見たようにVDの性能は実用レベルに達しており，登録操作の手間という点でもボタン操作に劣らない．しかし，多くの場面ではボタン操作と比較して明確な優位性がないのも事実である．また，人前で機械に話しかけることへの抵抗感もVD使用へのハードルとなる．

VD機能付き携帯電話機に関するモニタ調査[6]によると，VDを使い始めて最初のうちは，VD機能に満足しつつも人前で機械に話しかけることへの抵抗感も強いが，使い慣れるにしたがって抵抗感が薄れ，使用意欲が高まるようである．

こうしたことから考えて，VDは当面，これを使わざるを得ない人，および使うことによるメリットが大きい人から定着し，一般にも広まっていくのではなかろうか．たとえば，自動車内は有望な使用場所である．他人に聞かれる心配が少ないことも有利である．また，手が不自由であるとか視力が弱いなどの理由でボタン操作が困難な人にとっては，大きな助けになるはずである．最近の携帯電話は小型化が進んでボタンが小さくなり，高齢者にとっては操作しづらくなっている．これからの高齢化社会においては，VDの必要性はますます高まっていくものと考えられる．

現在商用化されているVDは，個人電話帳の音声版というべきものがほとんどであるが，Korea Telecomが試行提供しているボイスディレクトリ型のサービス[4]は，センタ型VDの新しい方向であろう．このサービスは，一般の電話帳のダイジェスト版に相当するものであり，個人電話帳と共存可能である．また，電話帳の保守・整備が個人の手に余る点，および認識処理量が多く端末には荷が重い点から，この種のサービスにはセンタ型がふさわしい．

最近，音声認識を組み込んだテレビゲームが発売され話題を呼んだ．このような玩具に慣れ親しんだ子供たちが大人になってボイスダイヤリングを違和感なく使いこなす時代がくることを期待したい．

参考文献
1) 河井, 大野, 中村: ボイスダイヤリングシステムの実装とフィールド試験, KDD R&D Report, No. 160, pp. 25-34 (1998).
2) 河井, 樋口: 電話網経由の連続数字音声の認識実験, 電子情報通信学会音声研究会資料, SP98-69, pp. 9-14 (1998).
3) Vysotsky, G. J. : VoiceDialingSM-The First Speech Recognition based Service Delivered to Customer's Home from the Telephone Network, Speech Communication, Vol.17, pp.235-247 (1995).
4) http://ktwww.kotel.co.kr/rndnews/news4_e.html
5) 山下, 金子, 木村, 塩野, 青木, 高橋: デジタル・ムーバN207HYPERの開発, NEC技報, Vol.52, No.2, pp.83-86 (1999).
6) 野村, 大蜘蛛: 携帯電話音声ダイヤルのモニタ評価, 日本音響学会講演論文集, pp.127-128 (Sep. 1998).

（平成11年5月7日受付）